Python for Data Science

Variance, Covariance, and Correlation

The difference between variance, covariance, and correlation is:

- *Variance* is a measure of variability from the mean
- *Covariance* is a measure of relationship between the variability (the variance) of 2 variables. This measure is scale dependent because it is not standardized.
- *Correlation/Correlation coefficient* is a measure of relationship between the variability (the variance) of 2 variables. This measure is standardized and is not scale dependent.

A more in-depth look into each of these is provided below. Here is data used to demonstrate the concepts on this page.

```
import os
import pandas as pd
import numpy as np

# Setting a seed so the example is reproducible
np.random.seed(4272018)

df = pd.DataFrame(np.random.randint(low= 0, high= 20, size= (5, 2)),
                  columns= ['Commercials Watched', 'Product Purchases'])

df
```

Commercials Watched	Product Purchase
14	7
12	15
6	7
7	1
5	10
x? = 8.8	? = 8.0
SD_x = 3.96	SD_y = 5.10

What is Variance?

Variance is a measure of how much the data varies from the mean. It can be represented as the following formula:

$$variance(s^2) = ?(x_n - x?)^2 / N-1$$

Where x_n= value point, x?= mean, and N= number of observations. The variance in the variable Commercials Watched can be calculated as:

x? (Commercials Watch) = (14 + 12 + 6 + 7 + 5)/ 5 = 8.8

variance(s^2) = ((14 - 8.8)^2 + (12 - 8.8)^2 + (6 - 8.8)^2 + (7 - 8.8)^2 + (5 - 8.8)^2) / (5 - 1)

variance(s^2) = 15.7

This can be calculated very simply in Python.

```
df['Commercials Watched'].var()
```

15.7

What is Covariance?

Covariance is a measure of relationship between 2 variables. It measures the degree of change in the variables, i.e. when one variable changes, will there be the same/a similar change in the other variable. The formula can be represented as:

covariance(x,y)= ?(x_n - x?)(y_n - ?)/N-1

It's the same formula as when calculating variance of a single variable, except instead of squaring each difference, we multiply by the other variables difference for that same observation. The covariance for Commercials Watched and Product Purchases can be calculated as:

? (Product Purchases)= (7 + 15 + 7 + 1 + 10)/ 5 = 8

covariance(x,y)= ((14-8.8)(7-8) + (12-8.8)(15-8) + (6-8.8)(7-8) + (7-8.8)(1-8) + (5-8.8)(10-8)) / 5-1 = 6.25

Again, this can be calculated easily in Python.

```
df[['Commercials Watched', 'Product Purchases']].cov()
```

6.25

The issue with covariance is that it requires both variables to be measured on the same scale to get some meaning, i.e. covariance is not standardized and it's interpretation is scale dependent. One cannot easily compare different covariance measurements unless the variables are all on the same scale. This is were correlation comes in.

What is Correlation and the Correlation Coefficient?

The measure of correlation overcomes the scale dependency of covariance by standardizing the measures. Standardizing measures makes it so the variables are on the same scale of measurement. The values are standardized which converts them into Z-scores. A Z-score is a standardized measure which measures variability from the mean.

The data are converted to Z-scores by taking the value point and subtracting the mean from it, and then dividing by the standard deviation. This can be represented in a formula as:

standardize value = (x - x?)/ SD_x

Where x_n= value point, x?= mean, and SD_x= standard deviation. To see what this look like, let's convert a few values. The standard deviation for Commercials Watched is 3.96, and the standard deviation for Product Purchases is 5.10.

```
Commercials Watched | Product Purchases
(14-8.8)/3.96 = 1.31 | (7-8)/5.10 = -0.20
(12-8.8)/3.96 = 0.81 | (15-8)/5.10 = 1.37
(6-8.8)/3.96 = -0.71 | (7-8)/5.10 = -0.20
```

One doesn't need to convert the data to Z-scores before performing the calculation as the standardization occurs in the correlation calculation. The correlation formula is as follows:

$$r(x,y)= ?(x_n - x?)(y_n - ?)/((N-1)SD_xSD_y$$

It's the same formula as the covariance formula, except we added an extra step when calculating the denominator. This equation is known as the Pearson correlation coefficient. Let's see what the formula looks like using the data.

$$r(x,y)= ((14-8.8)(7-8) + (12-8.8)(15-8) + (6-8.8)(7-8) + (7-8.8)(1-8) + (5-8.8)(10-8)) / (5-1)3.96*5.10 = 0.31$$

This can be calculated very easily with Python.

```
df[['Commercials Watched', 'Product Purchases']].corr()
```

0.31

A correlation coefficient will always be between -1 and 1. The closer the value is to -1 or 1, the strong the relationship, the closer to 0 then the weaker it is. If the correlation coefficient value is positive, it means as one variable increase so does the other, and if the correlation coefficient value is negative, it means as one variable increases the other decreases. If the correlation coefficient is negative, the way to see which variable increases/decreases is to plot the data. It's usually plotted as a scatter plot.

There are not set standards for what is considered a weak or strong correlation. It's usually field dependent, but a guideline is below.

r value	**Strength**
0.0 – 0.2	Weak correlation
0.3 – 0.6	Moderate correlation
0.7 – 1.0	Strong correlation

There are other equations to calculate correlation coefficients, such as Spearman's rank (a.k.a. Spearman's correlation), Kendall's tau, biserial, and point-biseral correlations. Each of which have different assumptions about the data that must be met in order for the calculations to be considered accurate. This page will not go into each of those as it's out of the scope of this page's goal, which is to explain what variance, covariance, and correlation are, the differences between them, and how they are very similar.

Loading...